JPRS: 3300

23 May 1960

SOVIET DEVELOPMENTS IN INFORMATION PROCESSING AND MACHINE TRANSLATION

by V. N. Toporov
and
V. I. Grigor'yev

U. S. JOINT PUBLICATIONS RESEARCH SERVICE
205 EAST 42ND STREET, SUITE 300
NEW YORK 17, N. Y.

F O R E W O R D

This publication was prepared under contract by the UNITED STATES JOINT PUBLICATIONS RESEARCH SERVICE, a federal government organization established to service the translation and research needs of the various government departments.

JPRS: 3300

CSO: 3901-D/17

SOVIET DEVELOPMENTS IN INFORMATION PROCESSING AND MACHINE TRANSLATION

FOREWORD

This translation series presents information from Soviet literature on developments in the following fields in information processing and machine translation: organization, storage and retrieval of information; coding; programming; character and pattern recognition; logical design of information and translation machines; linguistic analysis with machine translation application; mathematical and applied linguistics; machine translation studies. The series is published as an aid to U. S. Government research.

Previously issued JPRS reports on this subject include:

JPRS: 68, 241, 319, 355, 379, 387, 487, 621, 646, 662, 705, 729, 863, 893, 925, 991, 992, 1006, 1029, 1130, 1131, 1132, 1133 and 3225.

SOVIET DEVELOPMENTS IN INFORMATION PROCESSING
MACHINE TRANSLATION

[Following are translations of two articles taken from the Russian periodical Voprosy yazykoznaniya (Problems of Linguistics), Vol. VIII, No. 6, November-December 1959, Moscow.]

On the Introduction of Probability into Linguistics

(Below is the translation of the first article: V. N. Toporov, author, pages 28-35 of original periodical.]

The conscious application of certain elements of the theory of probability (initially as very elementary statistics) to research on language data is now more than 100 years old. However, the conscious utilization of statistical methods dates back more or less to the rise of linguistics (cf. G. Herdan, Language as choice and chance, Groningen, 1956, p. 1). For a variety of reasons it took a very long time before practical results were forthcoming. The sphere of application of probability methods to language analysis and, of course, the technique required were unclear to some scholars (even though they failed to recognize it). Others regarded language data purely as material to illustrate various methods of statistical analysis.

The failure of investigators seeking to use this approach was possibly due both to the absence of a genuinely scientific theory of language and to the fact that the exceptional complexity of the relations between the different linguistic elements at various language levels could scarcely be explained by this approach until mathematics itself felt the need of constructing a general theory of random processes to analyze the random values depending on one or more continuously changing parameters. It was apparently only such achievements as A. M. Lyapunov's central critical theorem of the probability theory, the study of the sequence of dependent random values ("Markov's chain"), the investigation of random processes in which the distribution of probabilities for the state of the system under study depends solely on the state already attained, and the new interpretation of the probability concept itself that created the preconditions for adequately validating the probability approach to language data. And now some maintain that these data are simply illustrative material for the theory of probability (cf. M. Boldrini, Le statistiche letterarie e i fonemi elementari nella poesia, Milano, 1948). But this view no more defines the contemporary situation than does the other extreme, which holds that aspects of the probability theory are useful only as a mathematical model to check some linguistic conclusion (cf. A. S. C.

Ross, "Philological probability problems," Journal of the Royal Statistical Society, Series B, Vol. XII, No. 1, 1950, also G. Herdan, op. cit., p. 6). However, regardless of one's view on the limitations of probability (particularly statistical) methods as applied to language analysis, it is necessary to bear in mind above all the special class of formal language structures capable of being explained on the basis of the law of large numbers ("macrolinguistic" structures) (B. Mandelbrot, "Linguistique statistique macroscopique" in the book by L. Apostel, B. Mandelbrot, and A. Morf, Logique, langage et theorie de l'information, Paris, 1957).

After the pioneering studies of Zipf and a pleiad of scholars during and after the war (Yule, Thompson, Ross, Herdan, Giraud, Mandelbrot, Fuchs, and others), the statistical approach was not only tolerated but regarded as being useful. In fact, even statistical research, which is concerned not with language ("langue") as such, but with its graphic representation, complicated by some facts not directly related to language (so-called literary statistics), made it possible to establish a series of statistical laws possessing profound linguistic meaning or permitting reformulation applicable to language proper.

Nevertheless, the investigations of recent years, it seems to us, have not fully clarified two aspects of the probability approach to language. First, although quite a few attempts have been made to consider the facts of language (the word "language" is not used here in its narrow technical sense) from the probability standpoint, linguists frequently continue to look upon it as having auxiliary value in all cases pertaining to linguistics. In any event, it is not generally realized that in a considerable number of instances the probability approach to language is the most suitable, if not the only one possible. Second, -- and this is true even of those who fully understand the advantages of introducing probability into the analysis of linguistic data -- only a few recognize the significance of probability (taken in the broadest sense) in determining the relationship between linguistic statements (so-called "laws") and language reality. The probability nature of this relationship is still unrecognized, nor is the modeling and operational role of linguistic concepts or statements clear. Linguists do not know what to do with the indeterminate residue in the form of facts that are not reflected in the model suggested by the given concept or statement. We must remember that "the conception of probability is not the instrument of some narrow scientific discipline; rather it is a fundamental conception on which is based a knowledge of reality and the interpretation of which determines the formulation of some theory of knowledge" (cf. H. Reichenbach, The theory of probability, Berkeley -- Los Angeles, 1949, p. 11; cf. the considerations he sets forth on p. 10).

With these preliminary remarks out of the way, we may now examine some problems arising in connection with the use of probability in linguistic research. Although the meaning of a linguistic theory or

the methods of describing an actual language are not completely clear in details, the question as a whole will apparently be settled depending on the problems embraced by the particular theory and the practical purposes in describing the language. We may thus speak about the independence or equivalence of different language descriptions only if their problems are different. However, there are conditions under which a comparative evaluation or two or more descriptions of the same language (or fragments of it) is possible if not essential. This applies primarily to two types of cases. First, when particular problems are the same, but the methods of solution are different and, moreover, not reducible to one another. Second, when it is a question of describing a language as a whole from the strictly linguistic point of view.

In either case the criteria of choice are purely logical: they provide for maximum completeness, self-sufficiency, noncontradictoriness, and simplicity of description (cf. H. J. Uldall, Outline of glossematics, Part 1, Copenhagen, 1957, pp. 20-35). However, there are often formidable difficulties involved in using them (particularly when a language is being described as a whole), since we still do not know for sure the kind of questions that an exhaustive description must answer. This uncertainty results, in turn, from the fact that we still do not know what probability limitations are imposed: (a) on a given actual language, (b) on human language in general, (c) on all semeiotic systems as a whole. In other words, these questions can be transformed into another kind of statement: we do not know what a particular language (e.g., Russian or English) is in relation to language in general. Theoretically we know only what Russian (or English) is in relation to Russian speech (or English). The other questions can also be reformulated in more or less the same way. In considering any of them, the relation of the inherent qualities considered therein to what is contrasted with them from without remains unclear.

Consequently we have an approximate idea of the areas (specific languages) into which language as a whole breaks down, but since the latter is not completely known to the investigator, he has no way of extracting the most essential information from the whole, the information which more or less determines the knowledge of the individual spheres into which language in general breaks down. Until problems of this kind are solved, any description of a language will be either incomplete or, to the contrary, uneconomical. Thus, whatever may be our understanding of a specific language, it presumes in essence a probability approach.

On the other hand, each language contains its own probability limitations imposed by one level of the hierarchical structure on the other. Identifying them would make it possible to ascertain the degree of determinateness of each level in relation to the neighboring one. The same thing is evidently possible with the approach to language analysis employed by the London linguistic school which assumes that language data are so to speak "structuralized" by the investigator

himself and grant him the right to begin his analysis at any point that he may select. Establishment of the principles and nature of the determinateness of one level (or fragment) of a language in relation to another and ascertaining the degree of interdependence of these levels should make it possible, one would think, to combine within a single general theory all three trends in modern structural linguistics, as discussed by K. L. Pike in his report read at the International Congress of Linguists in Oslo in 1957 (compartmentalization, abstraction, integration) (cf. K. L. Pike, "Interpretation of phonology, morphology and syntax," Reports for the Eighth International Congress of Linguists, Vol. II, Oslo, 1937, pp. 334-342).

In time there might well emerge the possibility of a new approach to the analysis of linguistic facts presupposing a high degree of formalization and the introduction of certain mathematical methods. Its objective would be to study the relations between the reflections of one level on another. The category of probability would undoubtedly play a fairly important part here. The practical value of this approach could scarcely be questioned. Among other things, it would demarcate the strictly phonetic phenomena from those used only morphologically, etc., which is one of the major tasks of linguistics on the synchronous as well as diachronic planes.

A major accomplishment of recent years in linguistics was the discovery of isomorphism between the planes of expression and content. However, the methods of detecting isomorphism and, in part, its forms are still, in our view, insufficiently precise. Thus two objections may be raised to the modern theory of isomorphism. First, the choice of facts both on the plane of expression and on the plane of content is such that in certain instances there is no guarantee against its being arbitrary (if only in part). Second, isomorphism of the elements within each individual plane is sometimes not determinable with sufficient distinctness; at other times it is uneconomical to operate with this concept in its present form, since even an isolated element has to be regarded as a compressed complex of elements (e.g., a syllable with a single sound in languages where syllabic structure is fairly complicated).

Such a view understandably arises not only in connection with the problem of isomorphism, but also when a language is described. In case of the latter the investigator is apparently confronted with two alternatives. He may re-establish the most complete scheme of possibilities (it is usually very logical and symmetrical) for a given category. This is how L. Hjelmslev treated case and how J. Kurylowicz handled, in part, aspect and tense (L. Hjelmslev, La categorie des cas, I-II, Aarhus, 1936; J. Kurylowicz, "Aspect et temps dans l'histoire du persan," Rocznik orientalistyczny, Vol. XVI (1950), 1953, and the same author's L'apophonie en indo-europeen, Wroclaw, 1956, p. 25 ff). Despite its complete logicality and simplicity, this method is not always the best, since it necessarily entails a knowledge of the most complete model possible by which other less complete schemes may be interpreted.

When it turns out, however, that there is a still more detailed variant, all the schemes based on the preceding maximum variant have to be reinterpreted. In addition, this approach is often quite impractical because of its redundancy (cf., for example, the analysis of the Bulgarian noun in terms of case relations), which tends to grow in the event of asymmetrical relations between the plane of expression and the plane of content.

On the other hand, it is natural and sometimes even more convenient for the investigator to start not with a complete scheme of theoretically conceivable possibilities, but with their actual occurrence in a given language. The advantage of this method is that it facilitates the discovery of laws underlying particular typological schemes and developmental trends, whereas the first method of description is based essentially on the universality of categories without taking into account the fact that the very idea of universality is undermined by the existence of different types of phonological and grammatical categories. (The ideas concerned with the impossibility of unambiguous comparison of similarly named categories in different languages are clearly reflected in the studies of J. Firth and his school.) Thus when a language is analyzed on this plane, each element (including the categories) is revealed not as a rigorously determinate fact in an over-all scheme, but as one of the probabilities which may sometimes be realized in the same language and at other times not.

The whole problem of isomorphism could be reformulated in probability terms if it were a question of the common features in the principles of implication, when the presence of one phenomenon determines the presence or absence of another (this applies both to expression and to content). With this approach we could readily grasp the importance of setting the boundary of any element on a given level by the maximum growth of entropy (i.e., when the usual rules of implication cease to be observed) and of identifying the dominant links within the various elements by the direction and nature of entropy. It is evident that calculating the probability factors in an investigation of isomorphism would make the analysis more precise, since account would also be taken of those cases where there is no strict determinateness. The introduction of probability would make it possible to fashion new contacts between linguistics and information theory. It would also permit formulation of the problem of identity in terms of probability relations.

Another advantage of introducing probability is that it may sometimes make the search for invariants unnecessary. We know that in some instances determination of an invariant among the transformations in a language and its recovery from the variants is difficult, if not completely impossible. The fact is the discovery of invariant relations is complicated by the limitations (linguistic and non-linguistic) imposed by any text.

Specifically, it often happens that the possibility of juxtaposing two phonemes or two morphological categories in a language model to all intents and purposes cannot be realized because of chance (as far as the particular phenomenon is concerned) coincidence, and the scholars are compelled to set up complex systems of indirect oppositions. The usefulness of introducing probability is obvious if only because it does not demand of a language more than it can give in a compressed form, which, strictly speaking, is often unique and given to us directly. However, the problem of finding invariants and the related problem of identity frequently require linguistic experiments, the discovery of concealed potentialities, critical situations, and a unique reductio ad absurdum. It is clear that in many cases the probability picture is more useful from the practical point of view. However, this does not imply rejection of the idea of searching for invariants, which is central to the structural analysis of a language.

Note has already been taken of the desirability and economy of using probability analysis to ascertain the phonetic classification of sounds on spectrograms, without elucidating the problem of invariants, in the mechanical analysis of speech, decipherment, statistical investigation of aphasia (cf. G. Herdan, "Statistical interpretation of aphasia," Confinia Psychiatrica, Vol. I, No. 3, 1958), investigation of combinability of sounds in a syllable, and -- in more general form -- in solving the problem of linguistic expectancy and application to different planes. Probability methods have now been employed to establish the absolute chronology of certain linguistic facts, to determine the degree of linguistic affinity and extent of linguistic variety, and to make statistical studies of vocabulary and the distribution of various linguistic units (the probability character of semantic fields makes it possible in principle to study distribution here too). Some general theories of language and a statistical conception of style applicable apparently to different language levels and permitting a qualitative determination of style are bound up with probability analysis. This in turn is very helpful in solving the problem of identification of styles and considerably lightens the task of establishing the authorship of unknown texts if a minimum number of known texts of the authors are extant. Finally, there is the familiar probability approach to language from the standpoint of information theory as to a code with probability limitations.

There are some other ways of introducing probability into linguistics and allied disciplines. A new field has developed during the past decade in psychology -- "statistical behavioristics" -- which is a theory of stochastic (probability) processes applied to the sequence of responses in language (cf. G. A. Miller and F. C. Frick, "Statistical behaviouristics and sequence of responses," Psychological review, Vol. 56, 1949, pp. 311-324; J. B. Carroll, The study of language, Cambridge (Mass.), 1953, pp. 105-107). Characteristically, the study of so-called "transitional probability,"

i.e., the degree of probability that a given response will follow another given response or series of responses, unites this school with some purely linguistic schools which investigate on the probability plane what will be said provided that we know in advance the structural features of the situation in which communication takes place, and know further that the speaker is a member of a particular speech community (J. Firth). The essence of an act of speech communication and transmission of meaning from one person to another, as conceived by H. Walpole and developed by his successors, has apparently likewise found more rigorous expression in probability terms. (cf. H. R. Walpole, Semantics, Norton, 1941, p. 78 ff.; F. F. Nesbit, Language, meaning and reality, New York, 1955, p. 62 ff). The same approach seems to be feasible both in studying speech perception (the process of transformation of physical, undulating movements into linguistic units) [see Note below] and in solving the problem of translation not only from one language into another or within the same language, but also from one semeiotic system into another (on these types of translation, cf. R. Jakobson, "On linguistic aspects of translation" in the collection On translation, Cambridge (Mass.), 1959, pp. 232-239).
(Note: (1) Cf. D. B. Fry, Perception and recognition in speech, in the collection "For Roman Yakobson," the Hague, 1956, pp. 169-173.
(2) Although the author does not refer here directly to the probability approach, he appreciates the unquestioned value of a strictly deterministic view of the correlation between physical, undulating movements and the linguistic units based on them. Cf. his remarks on p. 170: "No theory requiring a single, unambiguous correlation between physical quantities and linguistic units can be easily reconciled with the experimental results that show wide variety in the physical keys capable of leading to recognition of an isolated speech sound.")

It will be noted, however, that many (if not most) of the above-mentioned applications of probability analysis belong primarily to speech ("parole") rather than to language ("langue") and are therefore sometimes marginal to the interests of those linguists concerned primarily with language structure, although in some cases reformulation of the various results in terms of language (but not of speech) is not too complicated. Moreover, an attempt was made quite recently to understand language (in the de Saussure sense) by adding to it the probability characteristics of its reflection in individual speech (G. Herdan). It is still too early to say whether this approach is feasible. (Among other things, one may suppose that the very distinction between language and speech and the creation of models of a concrete language already assumes a knowledge of certain statistical characteristics of the elements of these models.) However, it is essential now that we discover the uses of probability analysis that are directly applicable to language. The attention of scholars should be drawn initially to the types of probability processes capable of serving if only as a rough model of a language or of its individual fragments, to the probability characteristics of the indeterminate

(or seemingly so) residue of a language which is not taken into account in the constructed model of that language, and to the characteristics of the system of a language through its projection in concrete manifestations or, in the words of information theory, to analysis of a code through communication. (It will be borne in mind that in this case it is close to ergodic communications.) The introduction of probability into language analysis must be considered not an auxiliary method, but a necessity if for no other reason than that any language model will be only an approximate reflection of its structure. However, what is not included in a model or is only reflected in a summary and nondifferentiated way is to be found within the range of probability relations.

Just as every language system has an indeterminate residue along with a purely determinate part, so too on the diachronic plane we find a series of purely conditional and unambiguously explainable changes and a series of phenomena in two adjacent sections that does not lend itself to establishment of an unambiguous connection and yields only to probability analysis. When we reconstruct the prehistoric state of a language, the value of probability conclusions grows with the steady decrease of our knowledge about the determinate part.

There is no doubt that the introduction of probability indices with each reconstructed system or its individual parts, including forms, would represent some calculation of probabilities foreseeing all the responses possible in the particular theory. In essence it would be a question of reconstructing possible algorithms of the lost forms. In quite a few instances this probability reconstruction would not be economical (especially in the present state of the science), since we still do not know enough about linguistic topology nor have we gained sufficient familiarity with the probability rules of the play of implications at the various levels of language. However, even now the high degree of redundancy of potential probability reconstructions would be compensated for by the fact that we would immediately be confronted with many clear cases where the restoration of a given form or solution of the etymology of some specific word was generally pointless because of the lack of some reliable basic data, the virtually limitless number of solutions, or simply because the question was formulated incorrectly as far as the given linguistic fact is concerned.

Comparative and historical linguistics, especially classical Indo-European, exhibits one contradictory aspect; its use is sometimes helpful, at other times harmful, but always instructive. We are thinking here of what is usually called the law of regularity (or law of no exceptions) affecting phonetic changes. (Cf. especially the categorical formulation of this "law" in the foreword to H. Osthoff and K. Brugmann, Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen, Part 1, Leipzig, 1878.) This law was employed to uncover the set of correspondences on which the edifice

of modern Indo-European linguistics is erected. Whatever remained outside the set of correspondences seemed irregular and was explained as due to analogy, borrowing, later origin, loss of old forms, etc. It was gradually learned that there are languages related to Indo-European in which the indeterminate residue is no smaller than the determinate. Consequently, the effectiveness of the method based on the principle of regularity of sound laws as applied to these languages is not very great. Everything not provided for by the scheme could, of course, be given probability indices, but this would not be a sufficiently radical measure. The fact is that the phonetic laws themselves are essentially a kind of probability principles, a view that has now been confirmed by research on the mechanism and distribution of these laws (especially in linguistic geography). We are therefore dealing here not so much with a probability interpretation of a hitherto unexplained residue as with a new probability approach to the problem of reconstruction and comparison unlike the old, strictly deterministic approach.

The introduction of probability into comparative and historical linguistics shows that the barriers raised between genetic and typological investigations are very often artificial. In a typological analysis the determinate part is naturally still less and is itself determined by probability methods, the transition between any two determinate states being governed by the laws of probability. (Cf. R. Jakobson, "Typological studies and their contribution to historical linguistics," Reports for the Eighth International Congress of Linguists, Suppl., Oslo, 1957 and, in part, J. H. Greenberg, Essays in linguistics, Chicago, 1957.) Another field of application of these laws in typology is connected with the relationship between comparable systems in different languages, since even the same categories (e.g., voiceless and voiced consonants) play different parts (e.g., in Russian and some Caucasic languages).

The introduction of probability raises the extremely important -- and still unsolved -- question of predictability in language. Awareness of the fact that language has a tendency to greater efficiency (cf. the articles of O. Jespersen, concluding with "Efficiency in linguistic change" (Copenhagen, 1941); cf. also: J. Engels, "Y a-t-il du progres dans le langage?" Neophilologus, Jg. 40, aflev. 4, 1956) cannot as yet determine the specific ways of achieving this state. On the other hand, knowledge of the constituent parts of a language system, their interrelations, and characteristic configurations makes it possible in combination with statistical calculations to isolate in a given state not only the archaic but also the relatively new elements and thereby bring stratigraphic analysis to the temporal plane. The configuration of elements in the most recent layer obviously makes it possible, with some degree of probability, to determine the subsequent state, although not necessarily in a single variant. There are grounds for believing that prediction of the future state of a particular language is theoretically possible

(in this connection one should not ignore the polemic of C. Levi-Strauss and N. Viner, "Language and social laws," American Anthropologist, Vol. 53, No. 2, 1954; cf. in particular p. 157), as, for example, the reverse movement into the past without introducing any external data. (It hardly need be mentioned that we are referring here to prediction only of those elements which are determined by a system, not by factors lying outside language -- "langue.") However, the final solution depends on whether this future state is symmetrical with the past and whether the present state contains elements accessible to observation. Be that as it may, any description of the future state of a language would undoubtedly come down to some sets of probabilities.

Another aspect of the problem of predictability is based on the probability selection of elements of a given state with the elements determined by the preceding text. The outlook here is more promising and there is already a literature -- to be sure, it is not extensive or wholly linguistic -- dealing with expectancy in its various forms. Expectancy may also be expressed in the terms of linguistic time which is by nature topological and on which probability limitations have been imposed.

The subject of introducing the concept of probability into linguistics is of course not exhausted by the examples of possible applications mentioned above or by its future promise. Suffice it to say, however, that one of the most significant ways of converting linguistics into a science with rigorous research methods is to include probability. The introduction of probability into linguistics (and some other sciences) would impart an element of necessary precision and rigor and draw linguistics closer to some of the more exact sciences, while for practical purposes it would provide us with a simpler and more economical instrument for analysis.

On Code and Language

[Below is the translation of the second article: V. I. Grigor'yev, author, pages 128-130 of the original]

Modern information theory regards language as a code system. This approach makes it possible not only to apply to language the findings and principles of information theory, but also to create a basis for a more rigorous definition of many linguistic concepts. In the most general sense a code is a means of representing information in a form suitable for transmission through a channel of communication, the latter being understood as anything that serves as a transmitter of information, including the air between a speaker and a listener. Any code is a set of physically different signs each of which corresponds unambiguously to one out of many objects to which the action of the particular code extends. In the process of effecting communication the presence of a certain combination or certain sequence of objects to be coded determines the sequence in which the code signs are chosen by a person or device sending the communication. At the same time, this sequence of code signs sent through the channel of communication determines in turn the reverse process of selection of objects to be coded by the person or device receiving the communication. For example, the order of letters in a telegram determines the sequence of choice and transmission through the communication channel of the telegraphic code signs from which the receiving apparatus reproduces the text of the communication.

To ensure reliable communication, each code sign must be physically different from all the others. In a very simple situation, when there are few objects to be coded, it is possible to manage with only a few physical characteristics (parameters), with which the code signs are supplied. An example of this is the regulation of street traffic by lights, the code of which consists of three signs -- red, yellow and green lights. In most cases, however, the number of objects to be coded is so large that it is impossible to compile a code in which each sign would differ in physical parameter from all the others. That is why when many objects are to be coded, the required number of code signs are obtained by utilizing differences in physical parameters, particularly combination differences. Only a few elementary units are supplied with parametric differences and the code signs are prepared from these units by a set of different combinations. The base of the code is determined by the number of elements from which the code combinations are composed. In telegraphy and elsewhere, wide use is made of a code based on two elements, the so-called binary code. For example, Baudot's code frequently uses plus and minus pulses of the electric current as elements. All the signs in Baudot's code consist of the same number of pulses, hence the code is called uniform. Conversely, combinations of different length are used in nonuniform codes.

From the standpoint of information theory language is a non-uniform code with a large base. The signs of a language code consist of words, but many of the objects to be coded include things, concepts, and ideas represented by words. Phonemes are the elements of a language code. Like the elementary pulses of a telegraphic code, the phonemes in themselves do not correlate with the objects to be coded and therefore have no meaning. Phonemes are elementary code units differing from each other in physical parameters (differential characteristics) and used to obtain the required number of combination differences when the code signs (words) are composed.

Russian has some 40 phonemes and thus the Russian language code has a base of 40 elementary units. Since the length of a Russian word ranges from one phoneme to 20 or more, the maximum number of code combinations obtainable from Russian phonemes is virtually limitless. Actually, only a small number of the possible combinations is used to form the words. This means that Russian like any other language, uses the combination possibilities of phonemes in a very uneconomical way or, as expressed in the terms of information theory, the language is characterized by great redundancy in the use of its elements. This redundancy is rather useful. For example, a telegram can be correctly read because of it, even if some of the letters are transmitted erroneously. Redundancy ensures the high reliability of communication under the most difficult conditions of oral communication. This relation of code redundancy to degree of communication reliability is responsible for the wide employment in communications technology of special correcting codes in which a sign contains an admittedly superfluous number of elementary sendings, i.e., in which a letter may be coded not by five, but by seven or more pulses. Corrections are always possible when a correcting code is used because the sign has superfluous pulses, i.e., distortions occurring during the transmission process are overcome. Whereas in a correcting code redundancy is introduced in an efficient manner commensurate with the required degree of reliability, redundancy in a language develops spontaneously during its history and can be largely eliminated by a more economical coding of messages.

Phonemes are known to undergo sharp changes and they occur in speech in many combiner and free variants. We encounter this variance of elements not only in language but also in other code systems. The fact is combiner variance is not essential to the elements of a code. For example, when flags are used for signaling purposes by a fleet, the code elements -- the individual flags -- remain unchanged regardless of the way they may be combined with the other flags. In Morse code the elementary pulses are separated by pauses, and so here too the form of elementary pulse does not depend on the surroundings. The problem of combiner variance arises only when the code elements follow immediately after one another in such a way that the end of one element merges with the beginning of the next. This is particularly true of

Baudot's code in telegraphy. The nature of the current change with time in transmitting the combinations of Baudot's code can be represented as follows:

The figure shows that the plus pulse in this code has several combiner variants depending on which pulse -- the plus or minus -- precedes it and which follows, namely:

If combiner variants are not mandatory for the code elements, free changes in the elements are inherent in each code. Above all, depending on the characteristics of the apparatus and conditions of the communication channel, telegraphic pulses can be shortened or lengthened, the level may change, the wave tilt differ, etc. Despite this changeability of code elements, telegraphic communication is very stable. The reason is that despite all the changes, the basic physical difference between the code elements is preserved, in this case the difference in direction of the current. It is obvious that variance of phonemes should not remove their physical character. Despite all the changes, the phonemes retain the differential features whereby they contrast with one another. It is up to the linguist to make a strict determination of these differential features.

Combination differences play a decisive role in codes constructed on the combination principle, since it is precisely the combination differences that produce a high degree of variety in the signs. The combination differences themselves are a genuine physical fact which may be abstracted from the elements constituting the combination, just as the shape of a circle may be abstracted from a mass of circular objects. Comparing, for example, the following two sets of combinations

$\underline{\Delta}$	0	0		0	$\underline{\Delta}$	0
	+	+		+		+
0	1	1		1	0	1

we see that the combinations in each column are the same, despite the differences in elements, and, conversely, the combinations on one side are different despite the identity of their constituent elements. If the elements of the code are few and the code signs are largely based on combination differences, it is natural for the code structure to be

determined primarily by these combination differences. From the standpoint of code structure differences between the elements are purely of secondary importance. This fact is responsible for the relative independence of code structure from its constituent elements and the possibility of substituting some elements for others while preserving the structure of the code. In Baudot's telegraphic code, for example, we may use as code elements current and currentless pulses, pulses of sinusoidal oscillations of different frequency or different phase, etc., instead of sending of different direction.

In a language code the possibilities of substituting elements are extremely limited. Language is a natural code that develops in accordance with its own inner laws. The elements of a telegraphic code may of course be replaced with others, but this assumes the presence of an appropriate arrangement of the corresponding members and results in reconstruction or replacement of the apparatus. Language is a means of communication between members of large groups of people. The phonological system of a language took shape over a long period of development and so the replacement of all the phonemes with others is not feasible. Moreover, the role of phonemes in the structure of a language code is incomparably greater than the role of elementary pulses in the structure of a telegraphic code in which there are only two elements all told. Nevertheless, the large amount of redundancy in a language makes possible at least partial substitution of the code elements, which gives rise to dialectic and individual deviations from the phonological system, and changes in the phonological system during the process of language development.

The concept of a code is a very broad one. In essence, no information can be transmitted without first being coded in some way. Codes are employed both in the extremely simple case of traffic regulation by lights and in complex cases of signaling and transmission of messages through channels of communication. The activity of nerve cells and the transmission of biological traits by heredity is related to coding processes. Language occupies a special position in this extensive mass of phenomena. A language code is peculiar not only in the exceptional complexity of its structure, but also in the special relationship between the language sign and coded object, in the role of language in the life of society, in the patterns of its historical development. A knowledge of the common features and principles underlying the construction of codes may be useful to the linguist in his investigations of the specific qualities of language.

References

A. A. Kharkevich, Ocherki obshchey teorii svyazi [Outline of the general theory of communications], Moscow, 1955, Chapter 1.

G. Glison [Gleason], Vvedeniye v deskriptivnuyu lingvistiky [Introduction to descriptive linguistics], Moscow, 1959, Chapters XIX and XX.

5214

- END -